

# Variable subset selection via GA and information complexity in mixtures of Poisson and negative binomial regression models

**TYLER J. MASSARO**

*Department of Mathematics*

*University of Tennessee, Knoxville, 37996-0532, TN, USA*

*E-mail: massaro@math.utk.edu*

**HAMPARSUM BOZDOGAN**

*Department of Business Analytics & Statistics*

*University of Tennessee, Knoxville, 37996-0532, TN, USA*

*E-mail: bozdogan@utk.edu*

## Abstract

Count data, for example the number of observed cases of a disease in a city, often arise in the fields of healthcare analytics and epidemiology. In this paper, we consider performing regression on multivariate data in which our outcome is a count. Specifically, we derive log-likelihood functions for finite mixtures of regression models involving counts that come from a Poisson distribution, as well as a negative binomial distribution when the counts are significantly overdispersed. Within our proposed modeling framework, we carry out optimal component selection using the information criteria scores AIC, BIC, CAIC, and ICOMP. We demonstrate applications of our approach on simulated data, as well as on a real data set of HIV cases in Tennessee counties from the year 2010. Finally, using a genetic algorithm within our framework, we perform variable subset selection to determine the covariates that are most responsible for categorizing Tennessee counties. This leads to some interesting insights into the traits of counties that have high HIV counts.

## 1 Introduction

Count data often arise in healthcare and epidemiology data sets. For example, the number of outcomes observed (*i.e.* cases of a disease) in a specific group of people (*i.e.* citizens residing in Knox county). We tend to treat count data as realizations from a Poisson distribution, so that the probability of observing  $y$  events given an expected number of events,  $\mu$ , is precisely

$$\mathbb{P}[Y = y; \mu] = \frac{\mu^y \exp(-\mu)}{y!}. \quad (1)$$

In regression analysis, Poissonian data are most often related to a linear combination of independent variables (*i.e.*, covariates, exposures) via a log-link function. Hence, the Poisson regression model is defined as

$$\ln(\mathbb{E}[Y]) = \beta_0 + \sum_{j=1}^p \beta_j X_j, \quad (2)$$

where  $\beta_j$ ,  $j = 1, \dots, k$ , are the measured effects of  $k$  independent exposures,  $X$ . For a binary exposure,  $X_j$ , the amount  $\exp(\beta_j)$  is the increase (or decrease) in outcome incidence for subjects in which the exposure was observed, relative to subjects in which the exposure was not observed [8]. The amount  $\exp(\beta_0)$  is the expected number of counts in baseline subjects. Researchers are responsible for determining the criteria defining a baseline measurement.

A major assumption in Poisson regression modeling is that the mean and variance of the observed counts are equivalent. When this assumption is violated, as is often true of real data,

we say the data are overdispersed. A negative binomial distribution may be more appropriate than a Poisson for describing overdispersed count data, due to its excessive tail behavior [8].

The difference between negative binomial regression (also referred to as NB-2 models) and Poisson regression comes from our treatment of the observed count data. We still make the assumption that these counts come from a Poisson distribution, but we make the added assumption that the mean number of events,  $\mu$ , comes from a 2-parameter Gamma distribution.

## 1.1 Mixture models

Finite mixture (FM) modeling has emerged within the past 20 years as a popular means for handling unsupervised classification tasks. The underlying principle behind mixture modeling is that we treat our observed data as having been sampled from a convex sum of distributions [5, 15]. This concept is particularly useful for explaining overdispersion in count data as being due to an underlying heterogeneous population.

FM modeling has appeared in tandem with Poisson regression models. The extent of this type of research has mostly been limited to zero-inflated, zero-truncated, and hurdle-type models [8]. These models are designed to handle zero-counts, and are largely inappropriate for explaining heterogeneities due to covariates. Papastamoulis *et al.* (2014) described a methodology that can address covariate heterogeneity using FMs of Poisson regression models in the context of high-throughput sequencing data [11]. Their EM algorithms for estimating  $G > 10$  mixtures are freely available online from CRAN.

In addition, Park and Lord (2009) have developed FMs of Poisson and negative binomial regression models for explaining heterogeneities in vehicle crash data [12]. Of the modeling frameworks that have been suggested in the literature, theirs is most similar to ours in that they proposed using the information criteria AIC, BIC, and DIC to choose the optimal number of mixing components. However, these criteria did not give conclusive results, and so they were forced to subjectively choose a mixture of 2 NB-2 models. In the proceeding sections, we will discuss potential reasons for these inconclusive results. For more information on finite mixtures of negative binomial regression models, see also Zou *et al.* (2013) [17].

## 1.2 Model selection

Model selection exists as an alternative approach to classical hypothesis-drive statistics which require distributional assumptions and an arbitrary selection of a confidence level to evaluate  $p$ -values. Akaike was the first to propose a criterion for conducting model selection when he introduced his celebrated Akaike information criterion (AIC) in 1973. The formula, which involves a tradeoff between a candidate model's lack of fit given by the maximized log-likelihood function, and a penalty term for the number of parameters,  $n_k$ , is shown below:

$$\text{AIC} = -2 \log L(\hat{\theta}|X) + 2n_k. \quad (3)$$

When we compare 2 or more models that have been fit to the same set of data, we prefer to choose the model that minimizes the AIC score.

As it turns out, the original penalty term that Akaike proposed,  $2n_k$ , is not enough to prevent model overfitting [4]. Numerous other information criteria (IC) have been proposed, including but not limited to Schwarz's Bayesian Criterion (SBC or BIC) [14], the deviance information criterion (DIC) (which we will not consider here), the consistent form of AIC (CAIC) [4], and the information-theoretic measure of complexity (ICOMP) [5] which involves evaluating the inverse Fisher's information matrix (IFIM). The reader is directed to the respective citations for derivations of the formulae. We will point out that each of the criteria we consider in this paper all share the lack-of-fit term in common – what makes them different is how they penalize overparameterization.

## 2 Methods

### 2.1 Log-likelihood functions

Our ultimate goal when we carry out model selection in this framework is to determine the optimal number of components in a FM model, given a fixed maximum number of components (see [5] for heuristics on choosing this number). This optimal number is reflective of the number of subpopulations composing the entire population from which we have sampled our data. Before we can perform model selection, we must be able to generate the various criteria discussed previously. To do so requires that we compute the log-likelihood function.

Recall equation (2) when dealing with Poissonian data whose mean and variance are similar. It follows that  $\mathbb{E}[Y|X, \beta] = \exp(\beta^T X)$ . Since our outcome is sampled from a convex sum of Poisson distributions we have the following probability for observing  $n$ -dimensional count data,  $Y$ :

$$\mathbb{P}[Y; X, \beta, \pi] = \prod_{i=1}^n \sum_{g=1}^G \pi_g \frac{\exp(y_i \beta_g^T x_i) \exp(-\exp(\beta_g^T x_i))}{y_i!}. \quad (4)$$

In equation (4),  $G$  is the total number of components in the FM model;  $\pi_g$  is the weight assigned to component  $g$ ; and,  $\beta_g$  refers to the effects in the  $g$ -th component, as this will vary within components. It follows directly from equation (2) that the log-likelihood is

$$\log(L(\beta, \pi; Y, X)) = \sum_{i=1}^n \left\{ \sum_{g=1}^G [\log(\pi_g) + y_i \beta_g^T x_i + \exp(-\beta_g^T x_i) - \log(y_i!)] \right\}. \quad (5)$$

When data are significantly overdispersed, recall that we assume our  $n$  counts,  $Y$ , come from a Poisson distribution whose mean parameter,  $\lambda$ , is Gamma-distributed. This additional constraint gives us the following probability density for  $Y$ :

$$\mathbb{P}[Y; X, \beta, \pi, \alpha] = \prod_{i=1}^n \sum_{g=1}^G \left\{ \pi_g \frac{\Gamma(y_i + 1/\alpha_g)}{\Gamma(1/\alpha_g) y_i!} \left( \frac{\alpha_g \beta_g^T x_i}{1 + \alpha_g \beta_g^T x_i} \right)^{y_i} \left( \frac{1}{1 + \alpha_g \beta_g^T x_i} \right)^{\frac{1}{\alpha_g}} \right\}. \quad (6)$$

The interpretation of  $\pi$  and  $\beta$  follows as before. The  $\alpha_i$  are overdispersion parameters, satisfying  $\text{var}(Y) = (1 + \alpha \beta^T X) \beta^T X$  (note:  $\text{mean}(Y) = \beta^T X$ ). The quadratic dependence of the variance on the mean is why we refer to these as NB-2 models.

Finally, following directly from equation (6), the log-likelihood function for the negative binomial regression model is shown below [3]:

$$\begin{aligned} \log(L(\beta, \pi, \alpha; Y, X)) &= \sum_{i=1}^n \left\{ \sum_{g=1}^G \left[ \log \left( \Gamma \left( y + \frac{1}{\alpha_g} \right) \right) - \log \left( \Gamma \left( \frac{1}{\alpha_g} \right) \right) + y_i (\log(\alpha_g \beta_g^T x_i) \right. \right. \\ &\quad \left. \left. - \log(1 + \alpha_g \beta_g^T x_i)) - \left( \frac{1}{\alpha_g} \right) \log(1 + \alpha_g \beta_g^T x_i) - \log(y_i!) \right] \right\}. \end{aligned} \quad (7)$$

Note that when  $\alpha = 0$ , the data are not overdispersed and so the assumption that  $\text{mean}(Y) = \text{var}(Y)$  is satisfied. Using our methodology, if the overdispersion parameter is below a certain threshold, we elect to perform Poisson regression, otherwise we perform NB-2 regression. This is especially important for computing the log-likelihood functions once we carry out scoring.

### 2.2 Initialization

The choice of an appropriate initialization scheme was a source of great concern for Papastamoulis *et al.* in their algorithm [11]. They used a ‘‘Small-EM’’ strategy from Biernacki *et al.* (2003), which was necessary for dealing with a large number of component mixtures ( $G \gg 5$ ) [2].

We are not interested in exploring the possibility of more than  $G = 5$  components unless there is some justifiable epidemiological reason for doing so. Because of this, we opted for using a Poisson mixture model to perform unsupervised classification of our observed counts (see [10]).

As a result, our methodology involves first sorting the count data into  $G \leq 5$  groups, and then assigning observations in each group to an initial Poisson or NB-2 regression model.

### 3 Results

All numerical simulations were carried out in MATLAB 2015a. We used ArcGIS 10.2 for generating the map of Tennessee counties.

#### 3.1 Simulated data

We begin using our proposed model framework to choose the optimal number of components in a simple set of simulated data. Each observation in the data is an ordered pair,  $(x, y)$ , where  $x \sim N(\mu_i, 1)$ , and  $y \sim \text{Poisson}(\lambda_i)$ ,  $i = 1, \dots, 4$ . The parameters  $\mu_i$  and  $\lambda_i$  for all  $i$  are randomly generated integers between 1 and 50. In this way, each observation comes from 1 of 4 potential groups.

We tried fitting FMs of 1 to 5 Poisson regression models to these data; their scores are shown in Table 1. We see that all 3 scores select  $G = 2$  as the optimal number of components. The regression models corresponding to the solution when  $G = 2$  are shown in the right pane of Figure 1. The regression models each seem to do well modeling these data.

$G$	AIC	CAIC	ICOMP
1	2931.14	2939.74	2944.77
2	1392.98	1414.47	1416.78
3	1400.08	1434.46	1434.19
4	1428.51	1475.79	1472.49
5	1455.42	1515.59	—

Table 1: IC scores for mixtures of  $G = 1, \dots, 5$  Poisson regression models applied to the simulated data shown in Figure 1. All three of the scores are minimized for  $G = 2$ . Our regression framework failed to produce a stable IFIM when  $G = 5$ , so no ICOMP score was computed.

#### 3.2 Tennessee HIV counts

We now use real data taken from a 2014 county-level study on persons living with HIV in the U.S. South [9]. Readers are encouraged to refer to [9] for a summary of the entire data set. For demonstrating our methodology, we are interested only in the TN counties that were included in the study. These data are summarized in Table 4. For reference, the abbreviations we use when referring to individual covariates are summarized below in Table 2.

The count data are largely overdispersed ( $\text{mean}(\text{HIV}) = 166.45$ ,  $\text{var}(\text{HIV}) > 500,000$ ), so we proceed by using mixtures of NB-2 models. In Figure 2, we see that the model selected 4 regression models, which is indicative of 4 separate homogeneous subpopulations within Tennessee counties.

Given that there are 4 subpopulations, we used a genetic algorithm (GA) to select the subset of variables from the original 8 that contributes most to sorting the data. GA have been used numerous times for carrying out variable subset selection, although never for Poisson or NB-2 regression (see [1, 7, 16] for GA applied to other regression models). The results for performing GA subsetting in our model are summarized in Table 3. We see that variables 4 and 7 appear in each of the subsets most frequently chosen, and the model fitted with only variables 1, 3, 4, 5, and 7 is the best at minimizing the IC scores.

In the interest of parsimony, Table 5 shows coefficients from each of the 4 components in a FM model fitted with just the urban indicator and the non-Hispanic white proportion. We see that the proportion of whites in a county is significantly protective against HIV count in 3 out of 4 components, while the urban indicator is a significant risk for HIV count.

No.	ID	Description	Data type
–	HIV	Number of persons living with HIV	Continuous (count)
1	SCH	Proportion of persons with less than a HS education	Continuous
2	POV	Proportion of persons living below the poverty line	Continuous
3	INC	Natural logarithm of median income	Continuous
4	URB	Urbanicity indicator (population <50,000/>50,000)	Categorical (2)
5	UMP	Unemployment rate	Continuous
6	NHB	Proportion of Non-Hispanic Black persons	Continuous
7	NHW	Proportion of Non-Hispanic White persons	Continuous
8	HSP	Proportion of Hispanic persons	Continuous

Table 2: Description of the variables in the Tennessee HIV data. The outcome is HIV. Categorical variables have the number of levels listed in parentheses.

Subset	Rel. freq.	AIC	CAIC	SBC
1, 3, 4, 5, 7	17.00%	708.76	776.89	756.89
1, 2, 3, 4, 5, 7	10.00%	703.51	781.86	758.86
4, 7	10.50%	849.86	887.33	876.33

Table 3: Relative frequencies and IC scores for variable subsets chosen using the GA.

In components 1 and 4, urban indicator was not included in the regression analysis because all of the counties belonging to those components had urban indicators of 0 or 1, respectively. Based on Table 6, this makes sense. We see that component 1 contains counties with HIV counts less than 15, while component 4 contains counties with counts greater than 100. The proportion of individuals with a high school diploma or lower decreases in each component, as do the non-Hispanic white proportions. Interestingly, the poverty rate and unemployment rates also decrease. At the same time, the urban indicator, and the non-Hispanic black and Hispanic proportions increase with each component.

## 4 Discussion

Our results indicate that the proportion of white persons in a county seems to be a significant predictor of HIV count in that county. This is slightly inconsistent with the conclusions of Gray, *et al.* [9], who found that the proportion of black persons was the most significant contributor to HIV rate. That being said, our GA the non-Hispanic black variable was selected numerous times as a significant variable, however it was not one the most significant (results not shown).

That urban indicator was selected as a significant variable comes as no surprise. This result tells us that as the population increases, we should also expect to see an increase in HIV count. For numerous different reasons, this result makes perfect sense. Moreover, in Figure 3, we see that the counties with the highest HIV counts correspond to the some of the largest cities in Tennessee, including Knox (Knoxville), Davidson (Nashville), Hamilton (Chattanooga), and Shelby (Memphis).

It was also interesting to find that poverty rates and education levels tended to be negatively associated with HIV counts. Our best explanation is that, at least in Tennessee, population size is a confounder for education level and poverty rate. That is to say that as the population of a county increases, we also expect to see access to jobs and education increasing. It would be worth investigating these results further to determine what potential public health impacts exist.

## 5 Conclusion

We have proposed a novel framework for analyzing data sets in which the response variable is a count. Our approach allows us to systematically identify homogeneous subpopulations arising in the dataset, based on the covariates and their contributions to the count data in a Poisson regression model. We are also able to select the covariates that contribute most to determining the aforementioned subpopulations. Using this approach on a set of HIV count data in Tennessee has led to reasonable and quantifiable results. This would suggest that we could apply this approach to other sets of data involving counts of other infectious diseases, such as influenza, dengue fever, or even the Ebola virus disease.

## References

- [1] Akbilgic, O. and H. Bozdogan (2011). Predictive Subset Selection using Regression Trees and RBF Neural Networks Hybridized with the Genetic Algorithm. *European Journal of Pure and Applied Mathematics* 4(4): 467 – 486.
- [2] Biernacki, C., C. Celeux, and G. Govaert (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(1): 561 – 575.
- [3] Biswas, S. (2013). `nbreg.m` – Negative Binomial Regression. *MATLAB Central File Exchange*. Accessed 2015.
- [4] Bozdogan, H. (1987). Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3): 345 – 370.
- [5] Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity, in *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, Vol. 2, H. Bozdogan (ed.). Kluwer Academic Publishers, Dordrecht, NED, 1994, pp. 69 – 113.
- [6] Bozdogan, H. (2000). Akaike’s information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*: 62 – 91.
- [7] Broadhurst, D., R. Goodacre, A. Jones, J. J. Rowland, and D. B. Kell (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta* 348: 71 – 86.
- [8] Dohoo, I., W. Martin, and H. Stryhn (2012). *Methods in epidemiologic research*. VER Inc, Charlottetown, PEI, CAN: 890 pages.
- [9] Gray, S., T. J. Massaro, I. Chen, C. J. Edholm, R. Grotheer, Y. Zheng, and H. H. Chang (2015). A county-level analysis of persons living with HIV in the southern United States. Submitted to *AIDS Care*, ID: AC-2015-01-0069. 15 pages.
- [10] Massaro, T. J., and H. Bozdogan (2015). On the selection of the number of mixtures in Poisson data using information complexity. Unpublished.
- [11] Papastamoulis, P., M.-L. Martin-Magniette, and C. Maugis-Rabusseau (2014). On the estimation of mixtures of Poisson regression models with large number of components. *Computational Statistics and Data Analysis*, in press: 10 pages.
- [12] Park, B.-J., and D. Lord (2009). Application of finite mixture models for vehicle crash data analysis. *Accident Analysis and Prevention*, 41: 683 – 691.

- [13] Paterlini, S., and T. Minerva (2000). Application of Genetic Algorithm and Neural Network in Forecasting with Good Date. In *Proceedings of the 6th WSEAS International Conference on RECENT Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing*: 19 – 27.
- [14] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461 – 464.
- [15] Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons Ltd., Chichester, GBR.
- [16] Vinterbo, S. and L. Ohno-Machado (1999). A genetic algorithm to select variables in logistic regression: Example in the domain of myocardial infarction. *Journal of the American Medical Informatics Association* 6: 984 – 988.
- [17] Zou, Y., Y. Zhang, and D. Lord (2013). Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis and Prevention* 50: 1042 – 1051.

## 6 Figures

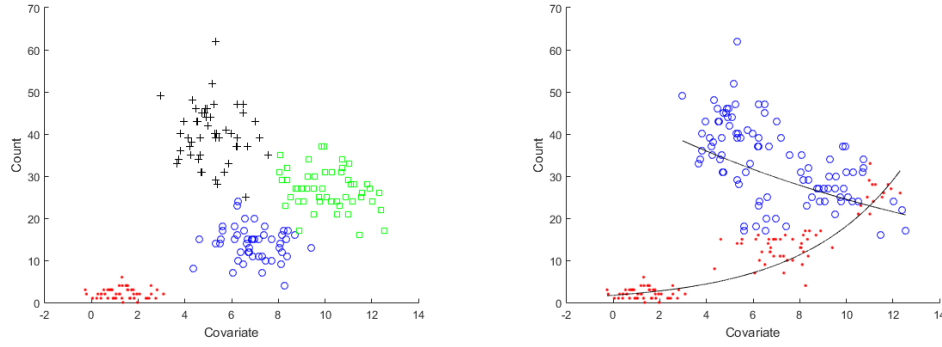


Figure 1: On the left, we see the data in its original form, with each observations coming from 1 of 4 distinct groups. On the right, we see the same data, with the mixture of two Poisson regression models that was chosen by ICOMP overlaid. Different colors and symbols indicate observations belonging to different mixtures.

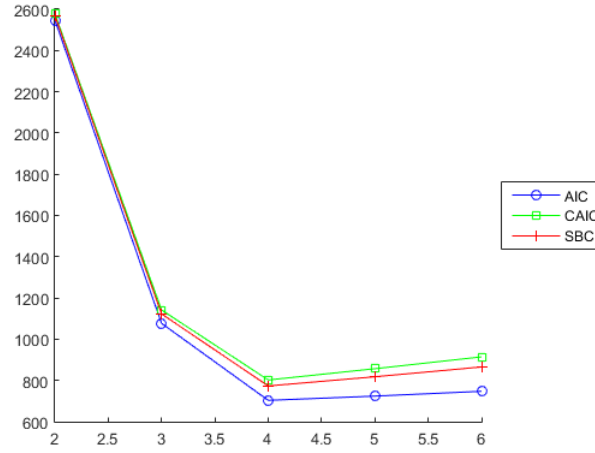


Figure 2: Information criteria scores for finite mixtures of  $G = 1, \dots, 5$  NB-2 models using the Tennessee county-level HIV data. We see that all 3 scores are minimized for  $G = 3$ .

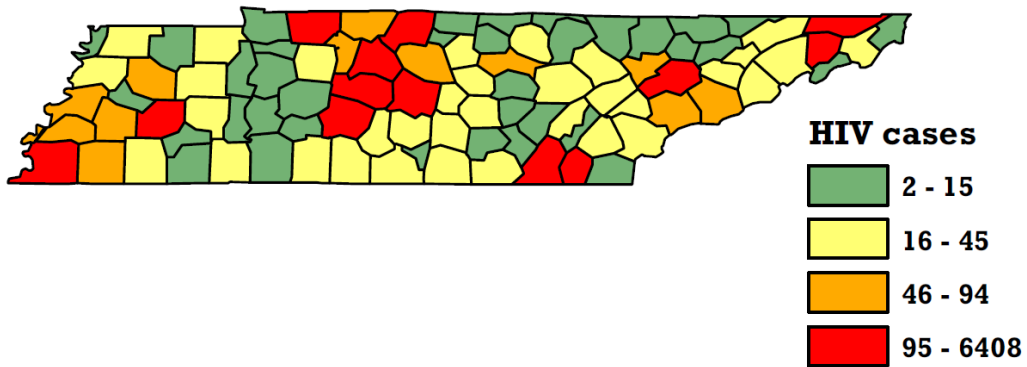


Figure 3: HIV cases by county in Tennessee.



## 7 Tables

Name	Mean	St. Dev.	Min	Max
HIV	166.45	770.43	2.00	6408.00
SCH	0.21	0.05	0.05	0.30
POV	0.17	0.04	0.05	0.30
INC	10.55	0.20	10.01	11.42
URB	0.31	0.46	0.00	1.00
UMP	0.11	0.03	0.05	0.18
NHB	0.07	0.10	0.00	0.52
NHW	0.88	0.12	0.39	0.99
HSP	0.03	0.02	0.00	0.11

Table 4: Summary statistics of Tennessee county-level HIV data taken from [9].

Variable	$\beta$	S. E.	2.5%	97.5%
Component 1				
URB	–	–	–	–
NHW	-2.14	1.30	-4.69	0.41
Int	4.06	1.20	1.70	6.41
Component 2				
URB	0.40	0.091	0.22	0.58
NHW	-1.85	0.51	-2.85	-0.85
Int	4.84	0.45	3.96	5.72
Component 3				
URB	0.35	0.11	0.12	0.57
NHW	-1.05	0.39	-1.81	-0.29
Int	4.94	0.28	4.38	5.50
Component 4				
URB	–	–	–	–
NHW	-7.12	1.40	-9.86	-4.37
Int	11.57	1.08	9.46	13.68

Table 5: Coefficients from the Poisson regression models in each component, using the variables selected by GA.

Variable	Mean	St. dev.	Min.	Max.
Component 1				
HIV	8.03	4.00	2.00	15.00
SCH	0.24	0.04	0.16	0.30
POV	0.20	0.04	0.10	0.30
INC	10.43	0.15	10.01	10.71
URB	0.00	0.00	0.00	0.00
UMP	0.12	0.03	0.07	0.17
NHB	0.03	0.05	0.00	0.26
NHW	0.93	0.06	0.68	0.99
HSP	0.02	0.02	0.00	0.09
Component 2				
HIV	27.55	8.89	15.00	45.00
SCH	0.21	0.03	0.15	0.30
POV	0.17	0.03	0.12	0.23
INC	10.54	0.10	10.30	10.81
URB	0.27	0.45	0.00	1.00
UMP	0.11	0.02	0.08	0.18
NHB	0.05	0.07	0.00	0.41
NHW	0.89	0.08	0.56	0.97
HSP	0.03	0.03	0.01	0.11
Component 3				
HIV	72.92	16.84	45.00	94.00
SCH	0.17	0.04	0.12	0.26
POV	0.14	0.05	0.09	0.23
INC	10.70	0.21	10.40	11.02
URB	0.54	0.52	0.00	1.00
UMP	0.10	0.02	0.07	0.15
NHB	0.14	0.15	0.01	0.49
NHW	0.81	0.15	0.45	0.94
HSP	0.03	0.01	0.02	0.06
Component 4				
HIV	1051.31	1913.88	101.00	6408.00
SCH	0.13	0.03	0.05	0.18
POV	0.13	0.03	0.05	0.17
INC	10.80	0.21	10.60	11.42
URB	1.00	0.00	1.00	1.00
UMP	0.09	0.02	0.05	0.12
NHB	0.16	0.15	0.02	0.52
NHW	0.75	0.16	0.39	0.94
HSP	0.05	0.02	0.01	0.10

Table 6: Summary statistics for counties belonging to each component.